
Visualizing dimensionality reduction of systems biology data

Andreas Lehrmann · Michael Huber ·
Aydin C. Polatkan · Albert Pritzkau ·
Kay Nieselt

Received: September 8, 2011 / Accepted: April 23, 2012

Abstract One of the challenges in analyzing high-dimensional expression data is the detection of important biological signals. A common approach is to apply a dimension reduction method, such as principal component analysis. Typically, after application of such a method the data is projected and visualized in the new coordinate system, using scatter plots or profile plots. These methods provide good results if the data have certain properties which become visible in the new coordinate system and which were hard to detect in the original coordinate system.

M. Huber was supported by the Deutsche Forschungsgemeinschaft (DFG) via a Heisenberg grant (Hu954/4) and a Heinz Maier-Leibnitz Prize grant (Hu954/5). A.C. Polatkan and A. Pritzkau were supported by the DFG Priority Program 1335 “Scalable Visual Analytics”.

A. Lehrmann
Center for Bioinformatics Tübingen, University of Tübingen
Sand 14, 72076 Tübingen
Tel.: ++49-7071-2970443, Fax: ++49-7071-295147
E-mail: lehrmann@informatik.uni-tuebingen.de

M. Huber
Wilhelm Schickard Institute for Computer Science, University of Tübingen
Sand 13, 72076 Tübingen
Tel.: ++49-7071-2977173, Fax: ++49-7071-295061
E-mail: michael.huber@uni-tuebingen.de

A.C. Polatkan
Center for Bioinformatics Tübingen, University of Tübingen
Sand 14, 72076 Tübingen
Tel.: ++49-7071-2970443, Fax: ++49-7071-295147
E-mail: polatkan@informatik.uni-tuebingen.de

A. Pritzkau
Bild- und Signalverarbeitung, University of Leipzig
Johannissgasse 26, 04009 Leipzig
E-mail: albert.pritzkau@medizin.uni-leipzig.de

K. Nieselt
Center for Bioinformatics Tübingen, University of Tübingen
Sand 14, 72076 Tübingen
Tel.: ++49-7071-2970443, Fax: ++49-7071-295147
E-mail: kay.nieselt@uni-tuebingen.de

Often however, the application of only one method does not suffice to capture all important signals. Therefore several methods addressing different aspects of the data need to be applied. We have developed a framework for linear and non-linear dimension reduction methods within our visual analytics pipeline SpRay. This includes measures that assist the interpretation of the factorization result. Different visualizations of these measures can be combined with functional annotations that support the interpretation of the results. We show an application to high-resolution time series microarray data in the antibiotic-producing organism *Streptomyces coelicolor* as well as to microarray data measuring expression of cells with normal karyotype and cells with trisomies of human chromosomes 13 and 21.

Keywords Dimension reduction · Principal Component Analysis · Independent Component Analysis · Local Linear Embedding · Systems Biology

PACS 87.18.Vf

Mathematics Subject Classification (2000) 62H25 · 15A18

1 Introduction

In biology as well as in many other scientific areas, high dimensional data sets arise in a natural way. In the case of gene expression data, this is mostly a consequence of the widespread utilization of DNA microarrays [29, 22] and the development of high throughput sequencing methods [30], especially during the course of the last decade. Both techniques allow the parallel analysis of thousands of genes and have thus led to an unprecedented flood of data. In order to find important signals in the large amount of data, a common approach is to combine feature extraction with subsequent dimensionality reduction. With a reduced dimensionality comes the advantage that the essential signals can easily be visualized either with the help of a parallel coordinates plot (PCP) as first introduced by Inselberg [15, 16] or, after a reduction to two or three dimensions, in a conventional two or three dimensional Cartesian coordinate system.

Among the vast variety of methods that are able to extract interesting signals out of large data sets are linear methods such as principal component analysis (PCA) [25, 10, 18] and independent component analysis (ICA) [14, 12, 13] as well as promising nonlinear procedures like the locally linear embedding (LLE) algorithm [26]. Although it is still a de facto standard, in our experience principal component analysis alone does not suffice to extract all important signals. Instead, it needs to be complemented with other methods in order to get the full picture [23]. For this reason, we have implemented several linear and nonlinear feature extraction methods including the ones mentioned above within our interactive and extendable visualization framework SpRay [6] that combines visual exploration with statistics-driven data analysis.

After visualizing the low dimensional representation of an originally high dimensional data set, a common problem concerns the interpretation of the obtained result. In this work, we put special focus on high dimensional data sets that are structured in the sense that the succession of features or experiments is not arbitrary but allows for a meaningful ordering. Examples are time series gene expression or similar systems biology experiments. If so, there is often substantial correlation between two adjacent features or experiments that transforms to

a similar location in n -dimensional Euclidean space. Visualization methods that reflect this inherent structure in order to support the user during the process of dimensionality reduction are therefore of great use. Here, we introduce two new visualization methods that we call ‘loading maps’ and ‘ l -neighborhood evolution matrices’. They assist the interpretation in case of the linear methods and give further insights into the structure of high dimensional data sets which help during parameter selection in the nonlinear case.

In an application study, we demonstrate the usefulness and potential of these two visualizations by applying them to a high resolution time series experiment conducted on the bacterium *Streptomyces coelicolor* [24]. We show that loading maps are a helpful tool for the identification of principal or independent components that encapsulate signals from genomic clusters with common functional annotation. Furthermore, we demonstrate the ability of l -neighborhood evolution matrices to detect ‘time clusters’ and significant biological events and to expose weak areas in strongly connected neighborhood graphs. The latter are an essential step not only in LLE but also in many other manifold learning algorithms (e.g., Isomap [32], Local Tangent Space Alignment (LTSA) [35] and Maximum Variance Unfolding (MVU) [34]).

In a second application, we conduct a principal component analysis on the samples of a microarray data set measuring expression of cells with normal karyotype and cells with trisomies. Here, the primary interest is to identify prevalent expression profiles among samples, regardless of individual genes expression patterns. We show that the loading map again provides a helpful overview for the comparison of the principal components allowing, in this case, to quickly identify significant samples.

Over time, a lot of effort has been put into solving the problems mentioned so far: A classic method to facilitate the interpretation of principal components is the varimax criterion developed by Kaiser [19]. For a visual impression of the dependency between the variables and the first two principal components, the circle of correlation [1] can be used. A novel approach using a so-called projection score to determine a suitable subset of variables that is easier to interpret was reported by Fontes and Soneson [7]. As far as LLE parameter selection is concerned, all available approaches are, to the best of our knowledge, of a purely computational nature. They range from hierarchical methods using residual variances [21] and methods using Spearman’s rho [20] to measures that assess the preservation of local geometry [33]. Our visual depiction of the neighborhood graph complements these mathematical methods in an excellent way. An interactive visualization environment was presented by Jeong *et al.* [17]. However, their framework is limited solely to the analysis of principal components and lacks the possibility of comparing results obtained by using different methods. The same is true for many programs that offer gene expression analysis, e.g., GeneSpring [2], TM4 [27] and Mayday [5].

In Section 2, we describe our mathematical framework, review linear and nonlinear feature extraction methods. Section 3 introduces our visualization methods in detail. Section 4 gives a short overview of SpRay and in Section 5 we present our application studies. We close with a discussion and an outlook.

2 Dimension Reduction Methods

2.1 Theoretical framework

Throughout the text, we assume the data to be represented by a matrix

$$X = (x_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} \in \mathbb{R}^{n \times p}, \quad (1)$$

in which the rows represent a set of features (e.g., expression values) and the columns represent a set of experiments (e.g., a time series), that is, x_{ij} is the value of a measurement of the i -th feature in the j -th experiment. When analyzing such a data set in a statistical environment, there are two perspectives that arise naturally: either we consider the features as variables and the experiments as observations or vice versa. To avoid case distinction, we will always assume in the following sections that the variables are given by the features. Subsequently, dimensionality reduction is done by a mapping

$$\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^{n'}, x_{\bullet j} \mapsto \phi(x_{\bullet j}) =: y_{\bullet j} \quad (2)$$

that maps every column vector $x_{\bullet j} \in \mathbb{R}^n$ of X to a new column vector $y_{\bullet j} \in \mathbb{R}^{n'}$, thereby defining a dimensionality reduced matrix

$$Y = (y_{ij})_{\substack{1 \leq i \leq n' \\ 1 \leq j \leq p}} \in \mathbb{R}^{n' \times p}. \quad (3)$$

If $\phi(x) = Ax$ for an $n' \times n$ matrix A (i.e., $Y = AX$), the dimensionality reduction method is called *linear*, and *nonlinear* otherwise. Often, ϕ is based on a certain cost function that one wants to maximize or minimize.

It is sometimes useful to think of a column of a matrix as being a realization of an underlying random vector. Where appropriate, we will switch between random vectors and their realizations in the form of a matrix without explicitly mentioning it. The same applies to statistical measures and their estimates.

Next, we give a short outline of the basic concepts of linear dimensionality reduction and then turn to a nonlinear approach.

2.2 Linear dimensionality reduction

Linear dimensionality reduction usually starts with centering of the data matrix X , a procedure that is equivalent to moving the coordinate system to the center of mass of the n -dimensional point cloud given by the columns of X . The coordinates with respect to the new coordinate system are now given by

$$X_c := X \cdot \left(I_p - \frac{1}{p} (1, \dots, 1)^T (1, \dots, 1) \right), \quad (4)$$

with \bullet^T being the transpose of a matrix or vector, and I_p being the p -dimensional identity matrix.

As pointed out earlier, we look for a matrix A , so that AX_c optimizes a method-specific cost function. We want to briefly review two popular cost functions that are based on the maximization of variance and independence and lead to a principal component analysis and an independent component analysis, respectively.

Principal Component Analysis Principal Component Analysis (PCA) [18] is a linear coordinate transformation that rotates a (centered) coordinate system in such a way that the coordinates with respect to the i -th axis have the maximum possible variance and are uncorrelated to the coordinates with respect to all axes $j < i$. Explicitly, we calculate the eigenvalue decomposition of the covariance matrix of X_c ,

$$\text{COV}[X_c] := (p-1)^{-1} X_c X_c^T = Q \Lambda Q^T, \quad (5)$$

and obtain the new coordinates satisfying the requirements as

$$Y_c = A X_c = Q^T X_c. \quad (6)$$

The matrix Q is the $n \times n$ matrix of the eigenvectors of the covariance matrix of X_c . All eigenvalue decompositions are assumed to be ordered in the sense that $\Lambda = \text{Diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_1 \geq \dots \geq \lambda_n$. The rows of Y_c are called *principal components* and for the new data matrix Y_c , we have $\text{COV}[Y_c] = \Lambda$ and $\text{tr}(\text{COV}[Y_c]) = \text{tr}(\text{COV}[X_c])$, i.e., the rows of Y_c are uncorrelated, $\mathbb{V}[y_{i\bullet}] = \lambda_i$ and the total variance has not changed.

The matrix Y_c is still of the same dimensionality as X_c . A common method to reduce dimensionality is to choose some value $\zeta \in [0, 1]$ specifying the amount of variance one wishes to maintain, to set

$$n' := \min \left\{ k \in \{1, \dots, n\} \left| \frac{\sum_{i=1}^k \lambda_i}{\text{tr}(\text{COV}(Y_c))} \geq \zeta \right. \right\} \quad (7)$$

and to limit the columns of Q in equation (6) to n' . This is based on the assumption that important signals contribute the most to the variance and are captured by the first few principal components.

Independent Component Analysis While independent random variables are always uncorrelated, the converse is generally not true. It is therefore a natural question to ask whether it is possible not only to decorrelate the data but also to maximize independence. This is indeed possible in several ways and Independent Component Analysis (ICA) [13] comprises a broad field of techniques which aims precisely at decorrelation and maximization of independence of the data at the same time. The basic principle is as follows: We assume that the centered data matrix X_c is the result of linear mixing of some underlying matrix Y_c whose rows are mutually independent and not normally distributed. The goal is to reconstruct this matrix Y_c by finding the inverse D of the (unknown) mixing matrix M and applying

$$Y_c := D X_c. \quad (8)$$

The ICA-model is greatly simplified by a whitening step that turns $X_c \in \mathbb{R}^{n \times p}$ into a new matrix $Z \in \mathbb{R}^{n \times p}$ with $\text{COV}[Z] = I_n$. This corresponds to decorrelation by PCA-preprocessing with subsequent scaling of the variances. If the eigenvalue decomposition $\text{COV}[X_c] = Q \Lambda Q^T$ has full rank, then

$$Z := \sqrt{\Lambda^{-1}} Q^T X_c = \sqrt{\Lambda^{-1}} Q^T M Y_c \quad (9)$$

is white. Note, that the dimensions of both matrices D and M are independent of the input data. If $\text{COV}[X_c]$ is rank-limited (as is the case whenever

$n \geq p$), the rows of Q^T and the rows/columns of A need to be limited to the rank $rk(\text{COV}[X_c])$. This is equivalent to a PCA-dimensionality reduction with subsequent scaling of the variances. Even in situations where it is not necessary, dimensionality reduction at this point is highly recommended. It is easy to see that we are now looking for the rows of an *orthogonal* demixing matrix $D' = (d'_{ij})$. A popular class of algorithms named FastICA to determine the rows of D' was introduced by Hyvärinen [14, 12] and is based on the maximization of non-Gaussianity. To maximize non-Gaussianity of the reconstruction $y_{i\bullet} := d'_{i\bullet}Z$, one can either maximize (approximations of) negentropy $J[d'_{i\bullet}Z]$ [11] or the absolute value of the fourth cumulant $|\kappa_4(d'_{i\bullet}Z)|$. The maximization of these values are computed in an iterative process involving gradient ascent or a fixed-point algorithm.

2.3 Nonlinear dimensionality reduction

Locally Linear Embedding The Locally Linear Embedding (LLE) [26] algorithm is based on the intuition that, although the columns of X are points in n -dimensional Euclidean space, they might in fact lie on a much lower dimensional submanifold of dimension $n' \ll n$. In this case, the data set can be represented in global, internal coordinates of the submanifold. To this end, every column vector $x_{\bullet j}$ of X is linearly approximated as well as possible by $\sum_{k=1}^p w_{jk} x_{\bullet k}$. Defining $W = (w_{jk})_{1 \leq j, k \leq p} \in \mathbb{R}^{p \times p}$ as the matrix of the weights, this approach leads to the minimization of the W -dependent total reconstruction error

$$K_1(W) = \sum_{j=1}^p \left\| x_{\bullet j} - \sum_{k=1}^p w_{jk} x_{\bullet k} \right\|^2. \quad (10)$$

To ensure locality and translational invariance, we enforce $w_{jk} = 0$ for all $x_{\bullet k} \notin N(x_{\bullet j})$ and $\sum_{k=1}^p w_{jk} = 1$ for all $1 \leq j \leq p$. Here, N is an arbitrary function that assigns each column vector of X a set of neighbors (e.g., based on their Euclidean distance). By definition, $x_{\bullet j}$ can not belong to its own neighborhood $N(x_{\bullet j})$. Note that the ‘size and form’ of the neighborhood of the individual points constitute the only free parameter of the algorithm. Therefore, its correct choice is of fundamental importance and crucially influences the result. We will come back to this important point shortly.

Because the optimal weights $w_{j\bullet}$ are invariant under translations, rotations and rescalings of $x_{\bullet j} \cup N(x_{\bullet j})$, they can, once found, be fixed, and in a second step, the cost function is minimized subject to the embedding, i.e., we now minimize

$$K_2(Y) = \sum_{j=1}^p \left\| y_{\bullet j} - \sum_{k=1}^p w_{jk} y_{\bullet k} \right\|^2. \quad (11)$$

This can be rewritten as

$$\sum_{i=1}^{n'} y_{i\bullet} \underbrace{(I_p - W)^T (I_p - W)}_{=: M} y_{i\bullet}^T = \text{tr}(YMY^T). \quad (12)$$

In order to avoid degenerate and non-unique solutions, the minimization is carried out subject to $\sum_{j=1}^p y_{\bullet j} = \mathbf{0}$ and $p^{-1}YY^T = I_{n'}$. Under these circumstances, for an eigenvalue decomposition $M = Q\Lambda Q^T$, the optimal embedding is given by

$$Y = \sqrt{p} \begin{pmatrix} q_{\bullet p-1}^T \\ \vdots \\ q_{\bullet p-n'}^T \end{pmatrix} \in \mathbb{R}^{n' \times p}. \quad (13)$$

3 Visualization

One of the biggest challenges in the areas of feature extraction and dimensionality reduction is to follow and understand the process from the input matrix X to the (possibly dimensionality reduced) output matrix Y . Although the underlying mathematical frameworks are well motivated, the application resembles a black box system, in which the choice of parameters and/or the interpretation of the results often is the major difficulty. We present two new visualization methods that assist the interpretation after feature extraction or dimensionality reduction and give further insights into the structure of high dimensional data sets by visualizing important intermediate steps. Although both techniques are applicable in a universal context, they are especially useful in situations where the features or experiments allow a meaningful ordering. This is the case in many real-world applications, for instance when analyzing a genome or conducting a time series.

Loading maps and their relatives While the mathematical elegance of a PCA is appealing, it has always been a major challenge to interpret the principal components. Because they are linear combinations of the original variables, it is a quite difficult task to assign a biological, physical or chemical meaning to them. To achieve more clarity, we propose to combine four theoretical measures that assist the interpretation with a visual component. Note that there are three fundamental terms in PCA: variables, observations and principal components. Usually, one is particularly interested in their mutual dependency [1], especially in the importance of

- a variable for a principal component (loadings):

$$L_0(x_{i\bullet}, y_{j\bullet}) = q_{ij}^2. \quad (14)$$

- a principal component for a variable (squared Pearson correlation):

$$L_1(y_{j\bullet}, x_{i\bullet}) = r(y_{j\bullet}, x_{i\bullet})^2. \quad (15)$$

- an observation for a principal component:

$$C_0(x_{\bullet j}, y_{i\bullet}) = \frac{\langle q_{\bullet i}, x_{\bullet j} \rangle^2}{\lambda_i (p-1)}. \quad (16)$$

C_0 describes the proportion of the variance of the i -th principal component that is caused by the j -th observation.

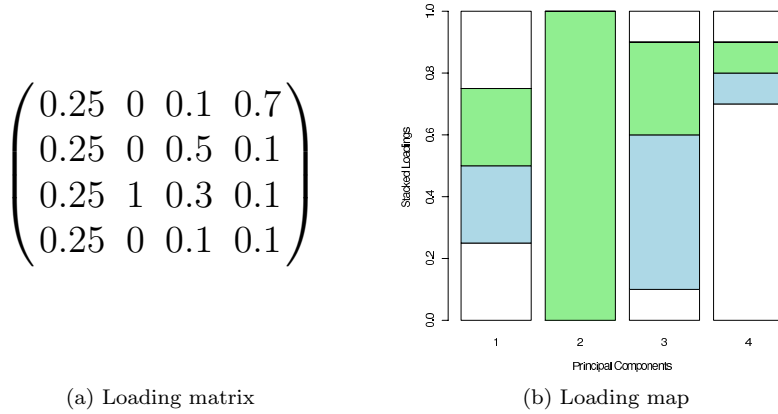


Fig. 1 Conversion of a loading matrix to a loading map. Figure 1(a) shows a hypothetical 4×4 loading matrix and Figure 1(b) the corresponding loading map. In this example, the variables X_2 and X_3 have been colored in blue and green, respectively, in order to show their positions in the map.

– a principal component for an observation:

$$C_1(y_{i\bullet}, x_{\bullet j}) = \frac{\langle q_{\bullet i}, x_{\bullet j} \rangle^2}{\|x_{\bullet j}\|^2}. \quad (17)$$

This is also known as the squared cosine and describes how much the i -th principal component contributes to the squared distance of the j -th observation.

Note that $\sum_{i=1}^n L_0(x_{i\bullet}, y_{j\bullet}) = 1$ for all principal components $y_{j\bullet}$. This is the basis for a visual representation that we term a *loading map*. For a particular principal component, a loading map displays the loadings of all variables as a stacked histogram, with the loadings sorted according to a given ordering. Because this type of representation is very space saving, it allows the parallel depiction of the loading maps of all principal components at once, thus giving within a single figure an immediate impression of the structure of the data set and the meaning of a particular principal component. Variables or groups of variables which are of special interest to the analysis (e.g., clusters of variables that are known beforehand) may then be colored to show their positions in the loading maps. After appropriate scaling, L_0 can be generalized to produce ICA-loading maps in a straightforward manner.

Consider the following illustrative example: We carry out a PCA on a $4 \times p$ matrix, $p > 4$, and collect all loadings in a loading matrix $L_0 = (L_0(x_{i\bullet}, y_{j\bullet}))_{1 \leq i, j \leq 4}$ that is subsequently converted to a loading map. Figure 1 demonstrates this process for a hypothetical loading matrix. Further down, we will give a detailed example where the variables are given by genes, the ordering is given by their sequence in the genome and the colors represent different functional clusters.

In fact, summation over the first component in any of these measures yields 1, so the same approach works for the other three measures (Eqs. 15-17) as well. This means we obtain a class of four visualizations that are very similar in appearance,

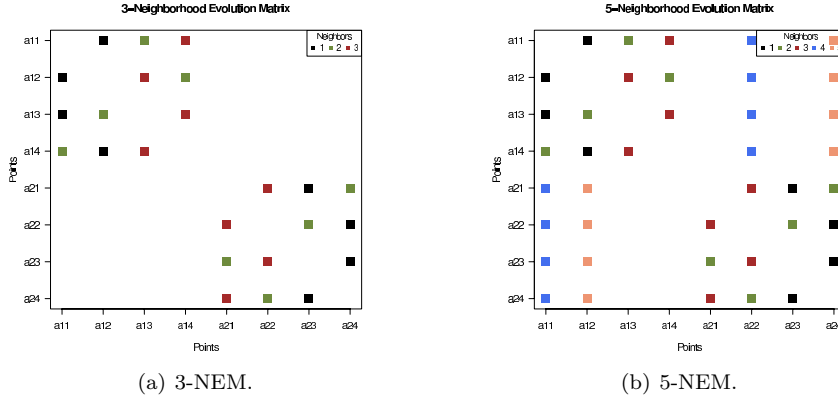


Fig. 2 l -neighborhood evolution matrices (l -NEMs). The figure shows a comparison between a 3-NEM (a) and a 5-NEM (b) originating from two clusters in n -dimensional space. While the 3-NEM shows a restriction of the neighborhood to the individual clusters, the 5-NEM indicates a ‘cluster overflow’.

yet they show completely different aspects of the analysis and complement each other so as to assist the interpretation of principal components.

l-neighborhood evolution matrices Our second contribution to the elucidation of important steps in the process of dimensionality reduction takes the form of what we termed *l-neighborhood evolution matrices* (l -NEMs). l -NEMs are a helpful extension to the use of classical dot plots, which have traditionally been used to compare two sequences, especially biological sequences like DNA or protein sequences. For l colors c_1, \dots, c_l , we extend this idea to the comparison of points in n -dimensional Euclidean space by coloring a cell (i, j) of a $p \times p$ grid with color c_k , if $x_{\bullet j} \in \mathbb{R}^n$ is the k -nearest neighbor of $x_{\bullet i} \in \mathbb{R}^n$. The similarity between two points can be determined by an arbitrary distance function, e.g., the Minkowski distance, correlation distance or any other distance function. If the distance function is symmetric and we consider all points as neighbors whose distance is smaller than a fixed radius R , then the neighborhood evolution matrix is symmetric as well (however not necessarily the colors of the dots). Note that this need not be the case for a fixed number of neighbors. The actual number of neighbors displayed is variable and allows for an adaptive view that shows how the neighborhood develops (hence the name l -NEM).

The coloring scheme is taken from ColorBrewer [9], which offers color sets to allow easy differentiation of the neighbors.

As an example, consider two compact clusters $A_1 = (a_{11}, \dots, a_{14})$ and $A_2 = (a_{21}, \dots, a_{24})$ in n -dimensional space (e.g., time points) whose single linkage distance is larger than any distance between two points within the clusters. In this case, we obtain a 3-NEM that is similar to that in Figure 2(a). The extension to a 5-NEM leads to a ‘cluster overflow’ and the neighborhood extends to the other cluster (see Figure 2(b)).

l -NEMs are of special use for confirming the number of neighbors that one intends to use in LLE. Originally, Saul and Roweis proposed increasing the neigh-

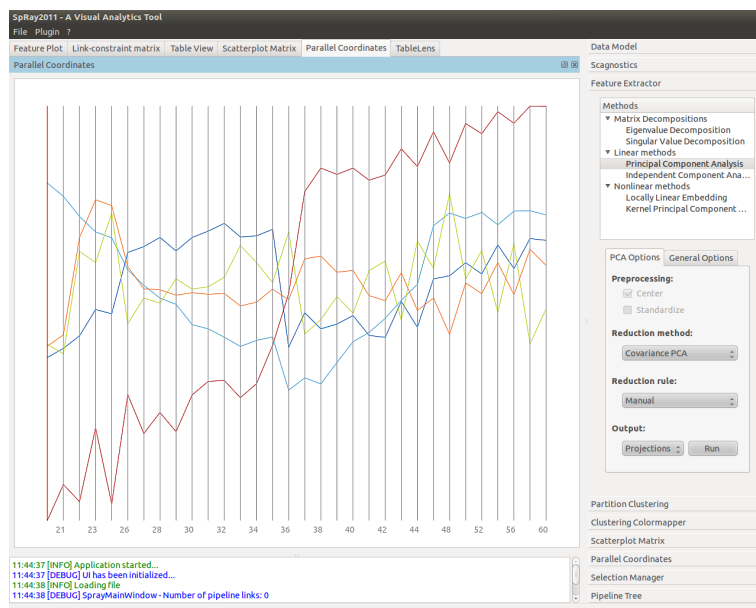


Fig. 3 The graphical user interface of SpRay. The options are accessible through a toolbox on the right, the visualizations through a tabbed navigation at the top of the main window.

neighborhood size until the neighborhood graph is strongly connected [28]. However, in situations like the one mentioned above, we may have strong connectivity only on account of there being a very small bottleneck. This is easily discovered using *l*-NEMs.

4 SpRay

All of the feature extraction/dimensionality reduction methods described so far, as well as the eigenvalue decomposition (EVD) and the singular value decomposition (SVD), have been implemented within our interactive visualization framework SpRay [6]. SpRay is a comprehensive, flexible and extendable software system that combines refined visual exploration with statistical methods to provide a unique visual analytics environment that proved to be particularly successful for the visual exploration of gene expression data. Different visual components like the parallel coordinates plot (PCP), the table lens and the scatterplot matrices are easily accessible in a linked view setting. SpRay uses a module-based approach and connects different processing modules to a pipeline that realizes the data flow. Along the pipeline, each module adds specific information to the data model and the visual components at the very end of the pipeline are able to access all of the information produced by the modules as and when required.

PCA, ICA and LLE are integrated into a single module named ‘Advanced Statistical Methods Library’ (ASML) that allows for fast switching between the different methods, thus permitting an immediate comparison of the results. The desired output dimension and the underlying mode of operation (i.e., reduction of

rows (features) or columns (experiments)) can be chosen in an interactive manner. Besides the basic procedures, the implementation includes several variants, extensions and modifications of the original algorithms. These include:

1. Principal Component Analysis: All four measures that assist the interpretation of the principal components (Eqs. 14-17); five different methods that recommend how many principal components should be kept; and a modification that uses correlation matrices instead of covariance matrices.
2. Independent Component Analysis: A cumulant-based and a negentropy-based approach to maximize non-Gaussianity; high and low pass filters to preprocess the data; and three contrast functions to approximate negentropy [12].
3. Locally Linear Embedding: Computation of strongly connected components in the neighborhood graph [31]; restriction to a specific submanifold in the disconnected case; and a version of the algorithm assuming X to be a distance matrix [28].

Figure 3 shows the graphical user interface of SpRay while analyzing a test data set. As can be seen from the selected options in the menu to the right, the parallel coordinates plot, in this case, shows the results after applying a reduction to five output dimensions with subsequent principal component analysis.

After processing the data, the loading maps as well as neighborhood evolution matrices can directly be generated on demand by the user within SpRay. For this, R is used, which is directly embedded in SpRay. In addition, SpRay writes the results obtained to a file in an R-compatible format. This technique allows for a seamless exchange between the two systems so that loading maps as well as neighborhood evolution matrices can be easily produced in R with adapted parameters.

SpRay is available for the community at <http://www-ps.informatik.uni-tuebingen.de/sprayWeb/>.

5 Application to expression data

To exemplify our visual analytics approach for dimension reduction we applied it to two different large-scale expression data sets. In the first example, reduction is applied to the genes, while in the second example, we demonstrate how the loading maps can be applied to a principal component analysis of experiments.

5.1 Time series expression data

For the first example, we use a data set from a large-scale transcriptomic study of the soil bacterium *Streptomyces coelicolor* grown under phosphate limited conditions [24]. From 20 hours after inoculation, samples were collected every hour up to 44 hours, and every two hours afterwards, resulting in a total of 32 time point samples. Global gene expression profiles were acquired for each sample using custom designed Affymetrix GeneChips [4]. The study revealed a very dynamic expression landscape as the organism undergoes a major metabolic switch from exponential growth to antibiotic production. We applied our dimension reduction methods to the matrix with 7893 variables (the genes) and 32 columns (the time points).

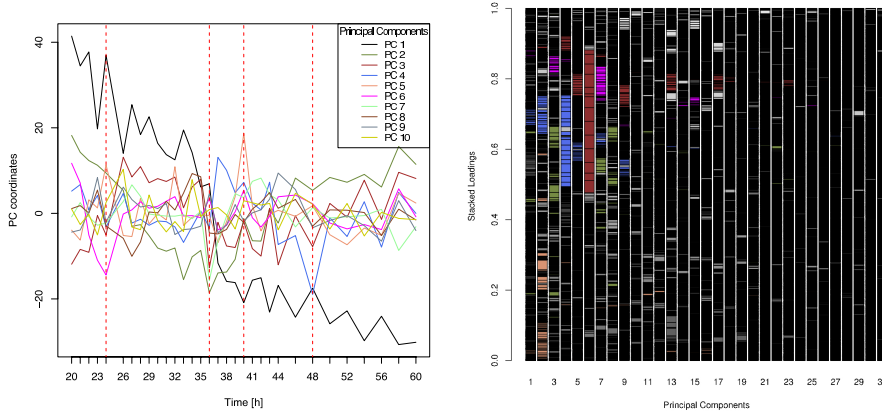
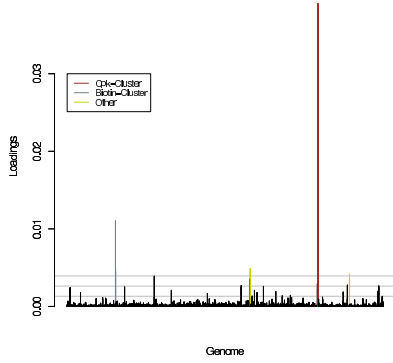


Fig. 4 Principal component analysis (PCA) and genomic loading map of a time series microarray experiment in the bacterium *S. coelicolor*. Here as a result, the profile plot of the first 10 principal components (left) is shown. The red dashed lines indicate four different important biological time points of expression changes. The genomic loading map on the right visualizes the PCA results in their genomic and functional context. For each component the stacked histogram of the loadings of all genes, that are sorted by their genomic location is shown. Selected gene groups with common functional annotation are color labelled. Red: genes of the Cpk cluster, blue: genes of the Act cluster, pink: genes of the Red cluster, green: phosphate dependent genes, orange: developmental genes.

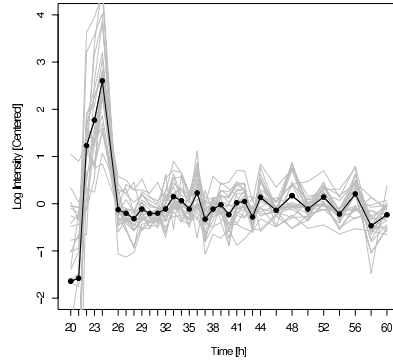
We first computed the eigenvalue spectrum which revealed that the eigenvalues of this time series matrix are rather homogeneous. This result attests to the large structural complexity of this data which has a variety of underlying signals. Figure 4 shows the parallel coordinate plot of the first ten principal components, capturing 90% of the data’s variance.

The profiles of the principal components reflect the dynamics of gene expression of this data with the most drastic change at 36 hours. In addition, at least three further time points, at 24, 40 and 48 hours, can be identified where at least one of the principal components shows a large increase or decrease in its coordinates. The authors in [24] have identified these time points to signal important onsets of regulation during the growth of the bacterium.

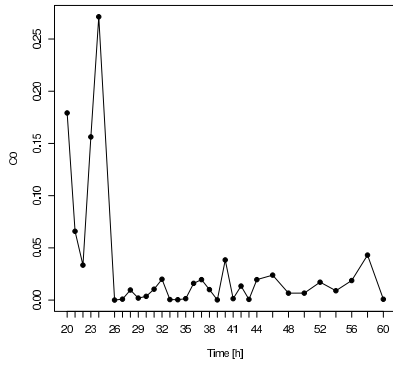
However, in order to gain greater insight in particular into the classes of genes that are responsible for the signals, we computed the four measures as introduced in Eqs. 14-17. We started with the analysis of the loadings (L_0 , Eq. 14) and their visualization by *genomic* loading maps (Figure 4). In a genomic loading map each principal component is visualized in terms of a stacked histogram of the loadings of each gene, with the genes ordered according to their chromosomal position. Since the loadings sum up to 1 for each PC, this visualization gives an immediate overview of which genomic locations are relevant for the PC. In addition, we colored genes with a common functional annotation. The visualization in Figure 4 clearly shows that no specific gene group dominates the first PC. PC 1 basically represents the signal of the global trend and it does not reveal local individual signals of specific functional relevance. The loading map also shows large loadings appearing



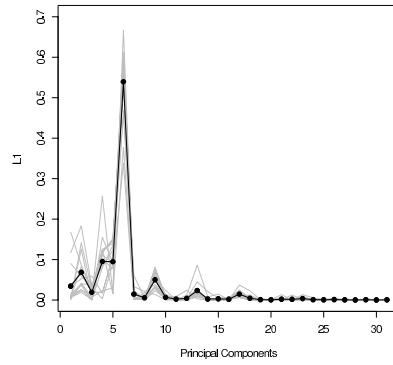
(a) Loadings (L_0 values) of genes along chromosome. Genes with loadings above the three-fold standard deviation are colored. One-fold, two-fold and three-fold standard deviations are denoted by grey horizontal lines.



(b) Expression profiles of genes with L_0 values above the three-fold standard deviation. The black profile is the mean profile of this group.



(c) Time dependence profile (C_0 values) of PC 6.



(d) L_1 values of genes of Cpk cluster (in grey). In black the mean L_1 values of this cluster are plotted.

Fig. 5 Complete characterization of the principal component 6 using the measures L_0 , C_0 and L_1 .

in chromosomal clusters, which clearly illustrates the common regulation of genes in the same genomic neighborhood, as is commonly observed in bacteria.

Loading maps offer a complete and at the same time compact view of all components, and therefore allow a direct comparison of all components. Genes with large loadings are clearly identifiable, while those with very small loadings are almost invisible.

In order to get an even clearer picture of individual components with high structural value, i.e., those that have few large loadings and most loadings close

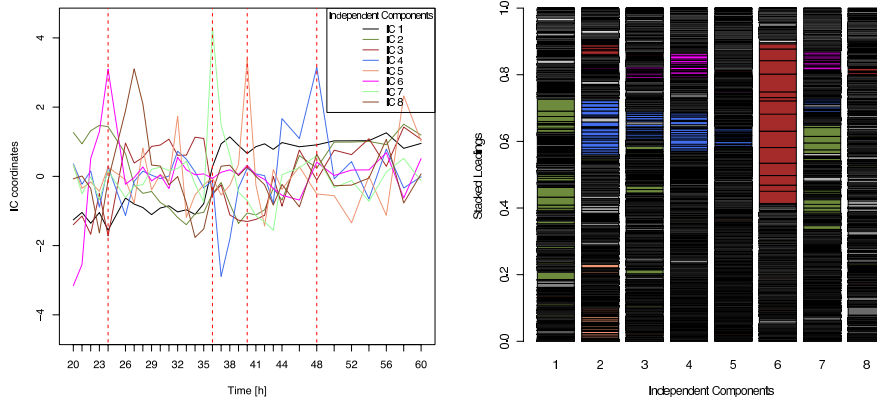


Fig. 6 Profile plot of independent components as a result of the independent component analysis (left) and genomic loading map (right). The same color map as in Figure 4 was applied.

to zero, we took a closer look at the sixth principal component. The visualization of its detailed characterization using the L_0 , C_0 and L_1 measures is depicted in Figure 5. From the visualization of the loadings of all genes in this PC, we find four clearly separated regions of genes, among them a group with very large loadings, whose corresponding profiles are very homogeneous, nicely showing the early up-regulation of these genes before 26h (Figure 5b). Furthermore, we see that PC 6 encapsulates this early expression event (Figure 5c) almost exclusively and that the *cpk* genes with very large loadings are also mostly captured by this PC (Figure 5d).

The application of independent component analysis to these data revealed the importance and complementarity of this method to PCA for this type of data. The profile plot of the first eight independent components (Figure 6 left) clearly shows the ability of the ICA to detect important signals by individual independent components (IC 6 for time point 24h, IC 7 for 36h, IC 5 for 40h, and IC 4 for 48h). In contrast to the first PC, the first IC represents the change from primary to secondary metabolism at 36h, marked by an almost homogeneous coloring of all those genes that have large loadings, as can be seen in the equivalent IC genomic loading map (Figure 6 right). Interestingly, the expression profiles of these genes are very similar to the profile of IC 1.

Finally, we applied LLE to this time series. To our knowledge the LLE method has so far not been applied to results from a time series expression experiment. In order to get a first impression of the neighborhood, we started by producing neighborhood evolution matrices for $l = 4$ and $l = 9$ neighbors (Figure 7). They clearly support our time pattern hypothesis, that already few neighbors suffice to detect a clear block structure, and where the respective time points are nicely related to important biological time points during the growth of the bacterium. In comparison to the 4-NEM, it is easy to see that the choice of 9 neighbors leads to a far better connected neighborhood graph because the four blocks (time clusters) start to interconnect. In Figure 7, this is indicated by dashed rectangles

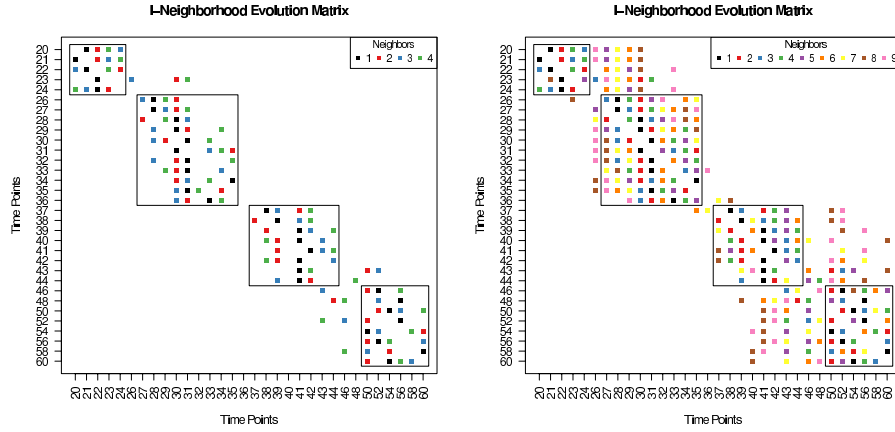


Fig. 7 Visualization of l -neighborhood evolution matrices of the expression time series in *S. coelicolor*. On the left $l = 4$, on the right $l = 9$ neighbors were chosen, using the Euclidean distance function. The colors were chosen with ColorBrewer [9].

that consist mainly of distant neighbors. Note also how distant neighbors tend to be offside the diagonal.

5.2 Trisomy Expression Profiling

In the second example, we used data from a genome-wide expression screen of cells with a normal karyotype and with trisomies of human chromosomes 13 and 21 [3]. Using Affymetrix microarray technology, gene expression in amniocytes (AC) from pregnancies with a normal karyotype and with trisomies of human chromosomes 13 and 21 was measured. For each of the three cell types, three replicates were used. From the normalized expression data we extracted all genes with a minimum variance of 0.7 across the experiments. We then conducted a principal component analysis on the matrix with 9 variables (the samples) and 1911 observations (the genes).

From the principal components we used the loadings (L_0 , Eq. 14) and created the *sample* loading map (Figure 8). In this loading map now each principal component is visualized in terms of a stacked histogram of the loadings of each sample, with the samples ordered according to their cell type. Furthermore, we also colored the samples with a common cell type. The visualization in Figure 8 clearly shows that about 80% of the loadings of the second up to the fifth principal component that are attributed to the trisomic cell types, dominated by the trisomy 21 cell type. This result supports the finding of Altug-Teber *et al.* [3], who showed that the trisomic samples can be clearly distinguished from the control samples.

6 Discussion

We have presented two new visualization methods, loading maps and l -neighborhood evolution matrices (l -NEMs), that assist linear and nonlinear feature extrac-

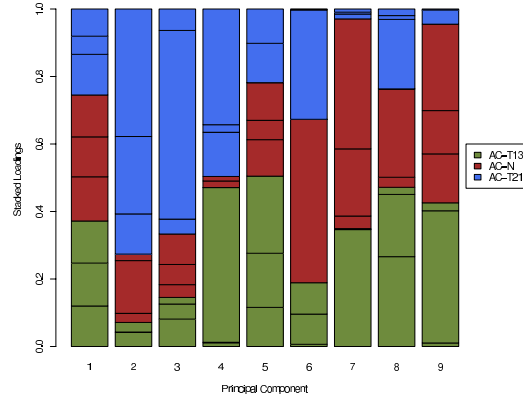


Fig. 8 Sample loading map of a microarray experiment comparing expression of human cells with autosomal trisomies with normal karyotype. The loading map visualizes the PCA results in their cell type context. For each component the stacked histogram of the loadings of all samples, that are sorted and grouped by their cell type is shown. AC-N: amniocyte cells with normal karyotype, AC-T13: amniocyte cells with trisomy 13, AC-T21: amniocyte cells with trisomy 21.

tion as well as an interactive framework, SpRay, that allows the visual exploration and comparison of many different dimensionality reduction methods. Our application to the study of the bacterium *Streptomyces coelicolor* has confirmed that both visualization techniques are of great value for detection of important features and experiments in large and complex data matrices. Using these methods, we were able to identify time points that go along with major metabolic changes and to extract clusters of genes that are responsible for these changes. We have seen that loading maps provide an effective second perspective that complements the parallel coordinates, facilitates their interpretation and explains their appearance. Furthermore, we could show that the application of only one method is often not enough by demonstrating that PCA alone was unable to isolate a vital event after 36 hours (the change from primary metabolism to secondary metabolism as a result of phosphate depletion). In contrast, the use of neighborhood evolution matrices indicated the importance of this time point at first glance. A more detailed investigation that included the application of ICA with subsequent production of ICA-loading maps was then able to identify the genes responsible for causing this event, namely phosphate-dependent genes and ribosomal genes. This demonstrates two things: First, that different methods are able to unveil different signals and, second, that the search for interesting signals is an interactive challenge/response procedure which makes it necessary to combine results, to rethink them in view of some new analysis and to see one result as a starting point for further steps.

In the second application we applied our loading map visualization to a comparison of different cell types. While in the first application example the genes were the variables, in this case the samples are the variables. The loading map of the samples allow conclusions whether certain sample types can be distinguished from others based on their expression profiles. Furthermore, we demonstrated that the stacked histogram is not focused on spatial relations.

In this work, we have so far restricted our examples concerning the practical applicability of loading maps and l -neighborhood evolution matrices to linear feature extraction and the locally linear embedding algorithm, respectively. However, both techniques are useful in a much broader sense and we want to give some examples of further applications.

l -NEMs give an excellent overview of the mutual location of points in n -dimensional space which easily identifies clusters, connections and distributions. This is especially the case if the points are ordered. These insights are of great use in all algorithms that depend on the construction of neighborhood graphs. The latter, in turn, are an essential part of many manifold learning algorithms including LLE, Isomap, LTSA and MVU in order to describe the local geometry around a point.

Loading maps can be generalized to other situations just as well. We already gave three more examples in the form of Eqs. 15-17. If we think of the measures as entries of a matrix, we see that this type of visualization is actually not limited to these four measures but works for every stochastic matrix. This includes all matrices that emerge from orthogonal matrices by squaring all of their entries. For the same reason, it is possible to combine our loading maps with other post-processing steps. For instance, the varimax criterion calculates an additional orthogonal transformation such that the sum of the variances of the loadings is maximized [19]. As the product of two orthogonal matrices is an orthogonal matrix, the new loadings obviously still fulfill the requirements for a loading map and can thus easily be visualized in the same way.

Regarding scalability of our implementation, a closer look at the feature extraction methods presented here reveals that all of them rely at some point on an eigenvalue or singular value decomposition. It is well known that both decompositions are computationally demanding tasks [8]. For biological data, however, we usually have data matrices $X \in \mathbb{R}^{n \times p}$ with $p \ll n$, a situation in which (under the agreement of a reduction of rows) the calculation of a thin SVD or the diagonalization of a $p \times p$ matrix is often feasible within a reasonable amount of time. For this reason, the computational complexity is in many relevant cases not an issue and our implementation scales well up to $2 \cdot 10^4$ rows, with the columns ranging from 10 – 500.

In some applications, it seems to be more appropriate to use slight variations of the visualization methods proposed. We therefore experimented with different modifications to the original loading maps and l -NEMs. For instance, it might sometimes be misleading that loadings belonging to a particular variable are, when viewed across several principal components, not necessarily at the same height. This is a consequence of the difference in the summed up heights of the preceding boxes of the stacked histogram. To eliminate this (not necessarily undesired) effect, one could instead consider representing the loadings through boxes of equal height, with the magnitude of the loadings visualized by a varying degree of transparency. This is, however, only feasible in the case where each box is visible, that is when the number of variables is not too large. When the variables represent all genes of a genome, as shown in our first application example, we consider our presented solution more appropriate.

Currently, l -NEMs differentiate between the neighbors of a particular data point through the use of a discrete set of colors. This is a plausible approach as long as the number of neighbors displayed is relatively small, say, not larger

than 9-12. This is also the maximum number of colors that ColorBrewer offers for a specific palette. Otherwise, it is going to get harder and harder to find an adequate set of colors that can be easily distinguished by the eye. For larger sizes of the neighborhood, we propose using either a linear RGB gradient between two colors or, again, a transparency-based encoding.

In conclusion, feature extraction followed by dimension reduction offers a helpful overview of high-dimensional systems biology data, such as expression data. We have shown, however, that no universal solution with automatic interpretation exists, rather analysis of the data should be understood as an interactive process. This process includes statistical and visual methods. With the loadings maps and the LLE evolution matrices, we introduced two new visualization techniques, that are highly suitable for the analysis of neighborhoods and loadings in the biological context, and therefore they are a valuable complement to the linear and non-linear feature extraction methods.

Acknowledgements We wish to thank Peter Wills for critically reading an earlier version of the manuscript. We also thank the anonymous referees for their thorough reviews and valuable comments that helped improving the presentation of the paper.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(4), 433–459 (2010)
2. Agilent Technologies: GeneSpring GX Manual. URL <http://www.chem.agilent.com/cag/bsp/products/gsgx/manuals/GeneSpring-manual.pdf>
3. Altug-Teber, Ö., Bonin, M., Walter, M., Mau-Holzmann, U.A., Dufke, A., Stappert, H., Tekesin, I., Heilbronner, H., Nieselt, K., Riess, O.: Specific transcriptional changes in human fetuses with autosomal trisomies. *Cytogenetic and Genome Research* **119**(3-4), 171–184 (2008)
4. Battke, F., Herbig, A., Wentzel, A., Jakobsen, Ø.M., Bonin, M., Hodgson, D.A., Wohlleben, W., Ellingsen, T.E., Nieselt, K.: A Technical Platform for Generating Reproducible Expression Data from *Streptomyces coelicolor* Batch Cultivations. In: H.R.R. Arabnia, Q.N. Tran (eds.) *Software Tools and Algorithms for Biological Systems, Advances in Experimental Medicine and Biology*, vol. 696, pp. 3–15. Springer (2011)
5. Battke, F., Symons, S., Nieselt, K.: Mayday - Integrative analytics for expression data. *BMC Bioinformatics* **11**(1), 121 (2010)
6. Dietzsch, J., Heinrich, J., Nieselt, K., Bartz, D.: Spray: A visual analytics approach for gene expression data. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2009)
7. Fontes, M., Soneson, C.: The projection score - an evaluation criterion for variable subset selection in PCA visualization. *BMC Bioinformatics* **12**(1), 307 (2011)
8. Golub, G.H., van Loan, C.F.: *Matrix Computations*, 1st edn. The John Hopkins University Press (1983)
9. Harrower, M., Brewer, C.: ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal* **40**(1), 27–37 (2003)
10. Hotelling, H.: Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**(7), 498–520 (1933)
11. Hyvaerinen, A.: New approximations of differential entropy for independent component analysis and projection pursuit. In: *Advances in Neural Information Processing Systems*, vol. 10, pp. 273–279 (1997)
12. Hyvaerinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks* **10**(3), 626–634 (1999)
13. Hyvaerinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*, 1st edn. *Adaptive and Learning Systems for Signal Processing, Communication, and Control*. Wiley-Interscience (2001)

14. Hyvaerinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural Computation* **9**(7), 1483–1492 (1997)
15. Inselberg, A.: The plane with parallel coordinates. *The Visual Computer* **1**(2), 69–91 (1985)
16. Inselberg, A.: *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer (2009)
17. Jeong, D.H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., Chang, R.: iPCA: An interactive system for PCA-based visual analytics. *Computer Graphics Forum* **28**(3), 767–774 (2009)
18. Jolliffe, I.: *Principal Component Analysis*, 2nd edn. Springer series in statistics. Springer (2002)
19. Kaiser, H.F.: The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**(3), 187–200 (1958)
20. Karbauskaitė, R., Kurasova, O., Dzemyda, G.: Selection of the number of neighbors of each data point for the locally linear embedding algorithm. *Information Technology and Control* **36**(4), 359–364 (2007)
21. Kouropteva, O., Okun, O., Pietikinen, M.: Selection of the optimal parameter value for the locally linear embedding algorithm. In: *Proceedings of the 1st international conference on fuzzy systems and knowledge discovery*, pp. 359–363 (2002)
22. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., Brown, E.L.: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**(13), 1675–1680 (1996)
23. Mannfolk, P., Wirestam, R., Nilsson, M., Sthlberg, F., Olsrud, J.: Dimensionality reduction of fMRI time series data using locally linear embedding. *Magnetic Resonance Materials in Physics, Biology and Medicine* **23**(5-6), 327–338 (2010)
24. Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, O., Sletta, H., Alam, M., Merlo, M., Moore, J., Omara, W., Morrissey, E., Juarez-Hermosillo, M., Rodriguez-Garcia, A., Nentwich, M., Thomas, L., Iqbal, M., Legaie, R., Gaze, W., Challis, G., Jansen, R., Dijkhuizen, L., Rand, D., Wild, D., Bonin, M., Reuther, J., Wohlleben, W., Smith, M., Burroughs, N., Martin, J., Hodgson, D., Takano, E., Breitling, R., Ellingsen, T., Wellington, E.: The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics* **11**(1) (2010)
25. Pearson, K.: On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**(6), 559–572 (1901)
26. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
27. Saeed, A., Bhagabati, N., Braisted, J., Liang, W., Sharov, V., Howe, E., Li, J., Thiagarajan, M., White, J., Quackenbush, J.: TM4 Microarray Software Suite. *Methods in Enzymology* **411**, 134–193 (2006)
28. Saul, L.K., Roweis, S.T.: Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research* **4**, 119–155 (2003)
29. Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235), 467–470 (1995)
30. Shendure, J., Ji, H.: Next-generation DNA sequencing. *Nature Biotechnology* **26**(10), 1135–1145 (2008)
31. Tarjan, R.: Depth-first search and linear graph algorithms. *SIAM Journal on Computing* **1**(2), 146–160 (1972)
32. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
33. Valencia-Aguirre, J., Ivarez Mesa, A., Daza-Santacoloma, G., Castellanos-Domnguez, G.: Automatic choice of the number of nearest neighbors in locally linear embedding. In: E. Bayro-Corrochano, J.O. Eklundh (eds.) *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science*, vol. 5856, pp. 77–84. Springer (2009)
34. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision* **70**(1), 77–90 (2006)
35. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing* **26**(1), 313–338 (2004)